



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

Heart Disease Prediction using Machine Learning

D.Vetriselvi¹, D. Arun Reddy², D. Shiva Reddy³, D. Narendra Babu⁴

Assistant Professor, Department of CSE Bharath Institute of Higher Education and Research, Selaiyur, Chennai,
Tamil Nadu, India

B. Tech Students, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu, India

ABSTRACT: Heart disease is a major cause of death worldwide, and early detection plays a crucial role in reducing mortality rates and improving patient care. This project focuses on building a heart disease prediction system using machine learning algorithms to assist in early diagnosis and clinical decision-making. The system uses patient health parameters such as age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, ECG results, maximum heart rate, exercise-induced angina, ST depression, slope of the ST segment, number of major vessels, and thalassemia. These features are used to train several classification algorithms, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The dataset used is taken from the UCI Machine Learning Repository. After data preprocessing and feature selection, models are evaluated based on accuracy, precision, recall, and F1-score. Among the tested models, Random Forest and SVM provided the highest accuracy, showing strong potential for real-world medical applications. The goal is to provide a simple and effective tool that can assist doctors in making faster, data-driven predictions. This project demonstrates how machine learning can significantly support the healthcare industry by automating disease prediction. Future work includes expanding the dataset, using deep learning methods, and developing a web-based interface for real-time access and usability.

I. INTRODUCTION

Heart disease, also known as cardiovascular disease, is one of the most common and deadly health problems across the globe. It refers to a range of conditions that affect the heart, including coronary artery disease, heart rhythm issues, and congenital heart defects. The rising number of cases and the cost of treatment make early diagnosis and preventive care critically important. Unfortunately, in many cases, heart disease goes undetected until it becomes severe. Therefore, it is essential to develop efficient systems that can help in predicting heart disease at an early stage using available medical data. With the rapid advancement of technology, the healthcare industry is now integrating artificial intelligence and machine learning techniques to improve diagnostic accuracy and treatment planning. Machine learning offers powerful tools that can learn patterns from historical data and make predictions or decisions without being explicitly programmed. In this project, machine learning algorithms are used to analyze a dataset of patients' health records to predict the presence or absence of heart disease.

The dataset includes various features such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, ECG results, maximum heart rate, exercise-induced angina, and more. These features are used as input to train and test different machine learning models including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The goal is to evaluate the performance of each algorithm and identify the one that provides the highest accuracy and reliability in prediction.

The developed model is not intended to replace medical professionals but to serve as a decision-support system that assists doctors and patients in identifying risk early. Such systems can be particularly beneficial in remote areas or in cases where regular check-ups are not possible. Ultimately, this project demonstrates the potential of machine learning in revolutionizing healthcare through early detection, improved decision-making, and personalized treatment.

II. LITERATURE REVIEW

The rapid advancement of machine learning techniques has brought significant improvements in the field of medical diagnostics, particularly in predicting chronic diseases such as heart disease. Heart disease remains one of the leading

causes of mortality worldwide, making early detection and prevention crucial for reducing health risks and improving patient outcomes. In recent years, researchers have increasingly turned to data-driven approaches to identify patterns and risk factors associated with cardiovascular conditions. Machine learning models, when trained on historical medical data, can uncover hidden correlations between various clinical attributes such as age, cholesterol levels, blood pressure, and lifestyle habits, thereby enabling accurate predictions.

1. Intelligent Diagnosis with Decision Trees

S. Palaniappan and Awang Norhashimah (2008) developed a Decision Support System (DSS) for heart disease diagnosis by applying data mining techniques on medical records. They used the C4.5 Decision Tree algorithm to extract meaningful patterns and rules from patient data, which could assist healthcare professionals in making early and accurate diagnoses. Their model effectively automated the decision-making process and demonstrated high predictive performance. However, they observed that while decision trees are interpretable, the model could be further improved by integrating ensemble methods to increase robustness and generalization.

2. Attribute-Based Heart Prediction

J. Singh and R. Kaur (2016) conducted a comparative study between Naïve Bayes and Decision Tree classifiers for predicting heart disease using a dataset with 14 clinical attributes. Their analysis revealed that the Decision Tree algorithm achieved higher accuracy compared to Naïve Bayes, highlighting its effectiveness in handling structured medical data. They emphasized the importance of selecting relevant attributes to enhance model accuracy and stability. Despite good results, the study acknowledged limitations in generalizing the findings to real-time clinical environments, pointing to the need for adaptive models.

3. Hybrid Classifier Models

M.A. Jabbar, Priti Chandra, and B.L. Deekshatulu (2017) proposed a hybrid classification approach that combined the strengths of K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). This model was tested on the Cleveland heart disease dataset and showed improved predictive performance over standalone algorithms. The hybrid model leveraged the local sensitivity of KNN and the global optimization capabilities of SVM, resulting in better classification accuracy.

4. Feature Optimization using Genetic Algorithms

M. Anbarasi et al. (2010) aimed to enhance prediction accuracy and reduce computational complexity by applying Genetic Algorithms (GA) for feature selection, followed by classification using the Naive Bayes algorithm. Their approach significantly reduced the number of attributes used in training while maintaining or even improving model accuracy. This not only reduced the computational cost but also improved model clarity. The study concluded that effective feature optimization is essential in medical data mining, although it requires careful domain expertise to avoid eliminating medically relevant features.

5. Comparative Study on Classifiers

U. Kumar and R. Sharma (2019) conducted a comparative study between Logistic Regression and Random Forest classifiers using the UCI Heart Disease dataset. Their results indicated that Random Forest provided higher prediction accuracy due to its ensemble nature, while Logistic Regression offered greater interpretability, making it easier for clinicians to understand. The study emphasized that the choice of algorithm should depend on the application's specific needs—whether the focus is on model transparency or maximizing predictive performance.

III. METHODOLOGY

Existing System

In the current medical landscape, the diagnosis of heart disease predominantly relies on manual interpretation by medical professionals based on a patient's medical history, physical examination, and diagnostic test results such as ECG, ECHO, stress tests, and angiography. Physicians often use their experience and clinical guidelines to identify the presence or risk of heart disease. However, this process is time-consuming, subject to human error, and dependent on the availability of experienced cardiologists. Additionally, in rural or under-resourced areas, there may be a lack of access to advanced diagnostic facilities, leading to delayed or inaccurate detection of heart-related issues. These

limitations emphasize the need for a more automated and reliable system that can support doctors in making timely and accurate decisions.

Some hospitals and diagnostic centres have begun integrating digital health records and rule-based systems to support diagnoses, but these are often rigid and unable to learn or adapt to new data patterns. Traditional statistical methods like regression analysis are occasionally employed for risk prediction, but they fail to capture the non-linear relationships between various health parameters. These methods also require domain-specific assumptions that may not hold true for all patients, making them less effective for large and diverse populations.

Furthermore, most existing systems do not utilize the vast amount of available patient data effectively. They lack intelligent algorithms capable of identifying hidden patterns or correlations among variables such as age, cholesterol, blood pressure, diabetes status, and more. As a result, these systems can be prone to underfitting, leading to generalized conclusions that may not reflect individual patient risk accurately. The inability to personalize predictions or update themselves based on new data makes them outdated and inefficient in today's data-driven world.

Security and interoperability are also challenges in the current systems. Many existing platforms operate in isolation and do not interact with other hospital systems or external databases. This fragmentation leads to data silos and restricts the ability of healthcare professionals to get a comprehensive view of a patient's medical history, which is essential for accurate heart disease prediction. Moreover, concerns about patient data privacy and lack of standardized protocols further limit the integration of AI-based tools in traditional medical practice.

Overall, while the existing systems provide a foundation for diagnosing heart disease, they are not optimized for early prediction, personalized assessment, or large-scale automation. There is a clear need for a modern, machine-learning-based solution that can analyze vast datasets, deliver accurate predictions, and serve as a decision-support tool for medical professionals across different regions.

Proposed system

The proposed system aims to address the limitations of traditional heart disease diagnosis by introducing a machine learning-based predictive model that can efficiently analyze patient data and predict the likelihood of heart disease with high accuracy. This system uses historical medical records and clinical data to train machine learning algorithms, which can then make accurate and timely predictions for new patients. By leveraging modern classification techniques such as Decision Trees, Random Forest, Naïve Bayes, and Support Vector Machines (SVM), the system can identify complex relationships between attributes like age, blood pressure, cholesterol level, chest pain type, fasting blood sugar, and more.

The core functionality of the proposed system is to serve as an intelligent decision support tool for healthcare professionals. It allows the entry or upload of patient details, processes them through trained models, and outputs a binary prediction—whether the patient is likely to have heart disease or not. The results are accompanied by confidence scores, which indicate the certainty of the prediction.

One of the key innovations in this system is the use of data preprocessing techniques, such as normalization, missing value treatment, and feature selection, which significantly enhance the model's accuracy. The system makes use of the Cleveland heart disease dataset, which contains multiple patient records with 14 important attributes. This data is split into training and testing sets to evaluate the performance of different algorithms and identify the one with the highest prediction accuracy. From the analysis, algorithms like Random Forest and Decision Tree have shown promising results due to their ability to handle missing values and multi-class classification.

In addition to prediction, the proposed system offers a user-friendly interface where users can input their health metrics manually and receive instant predictions. This interface is built with the aim of making the system usable not just by doctors but also by patients who wish to perform a preliminary self-assessment.

Overall, the proposed system offers a scalable, accurate, and efficient solution to assist healthcare providers in diagnosing heart disease at early stages. It not only improves diagnostic accuracy but also reduces the workload on

medical professionals, supports remote diagnosis, and enables proactive healthcare practices. The model's learning capability ensures it can continuously improve as more data becomes available, making it a robust tool for long-term deployment in hospitals, clinics, and telemedicine platforms.

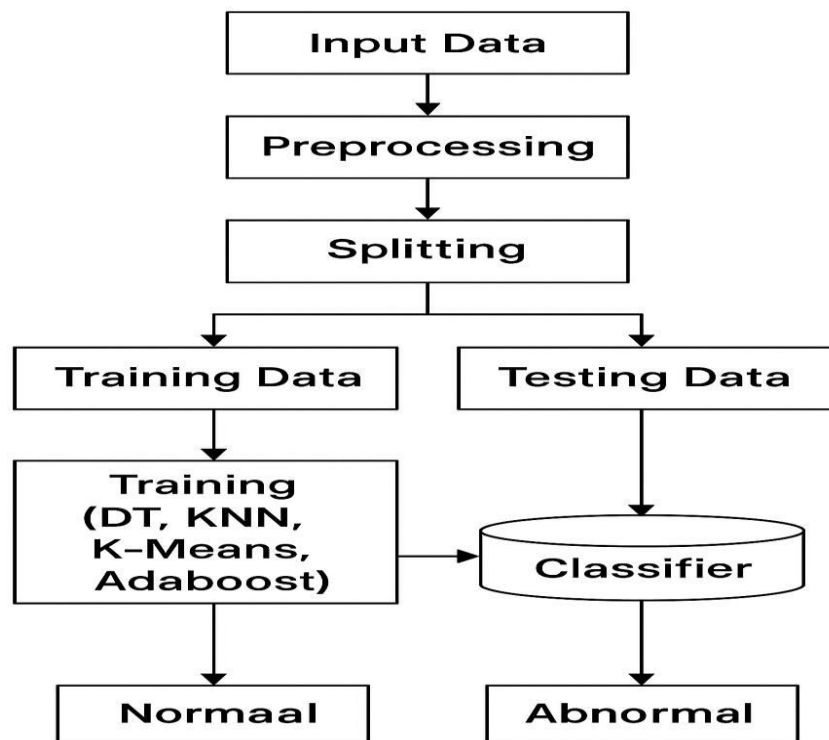


Fig 1: System Architecture

VI. IMPLEMENTATION PROCESS

1. Data Collection

The first and most critical step in the implementation is the collection of a reliable and relevant dataset. In this project, the UCI Heart Disease dataset is commonly used because of its accuracy and detailed clinical attributes. This dataset contains several important medical features such as age, sex, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, and more. Additionally, it includes a target attribute that indicates the presence (or absence) of heart disease in a patient. This labeled data is essential for training supervised learning models. Gathering this kind of structured and well-documented data ensures a solid foundation for the machine learning process.

2. Data Preprocessing

Raw data is often noisy and inconsistent, so data preprocessing is a vital step before training any machine learning models. The process begins by checking for missing or null values and handling them through imputation or removal to avoid model bias. Next, categorical variables such as chest pain type or thalassemia are converted into numerical form using techniques like label encoding or one-hot encoding. Additionally, feature scaling (normalization or standardization) is applied to numerical attributes to bring them to a uniform scale, which improves the efficiency of algorithms like SVM and KNN. Finally, the dataset is split into training and testing subsets, usually with an 80:20 ratio. This ensures the model is trained on a large portion of the data while still reserving a separate portion for unbiased evaluation.

3. Model Selection and Training

Once the data is clean and structured, the next step is to select appropriate machine learning algorithms to build prediction models. In this project, five classification algorithms were chosen: Support Vector Machine (SVM), Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), and Decision Tree. Each algorithm is implemented using Python's Scikit-learn library. These models are trained using the training dataset, during which they learn to identify patterns and relationships between the input features and the target outcome. Hyperparameters may be tuned using grid search or cross-validation to optimize the performance of each model. The purpose of using multiple algorithms is to compare their strengths and weaknesses for heart disease prediction and identify the most accurate and reliable model.

4. Model Evaluation

After training, the models are evaluated using the testing dataset to measure how well they generalize to unseen data. Several performance metrics are used for evaluation, including accuracy (how many predictions were correct), precision (how many positive predictions were actually positive), recall (how many actual positives were correctly predicted), and F1-score (the harmonic mean of precision and recall). A confusion matrix is also generated for each model to visualize true positives, true negatives, false positives, and false negatives. This comparison helps determine which model performs best for this specific problem. The model with the highest accuracy and balanced precision-recall trade-off is chosen for deployment in the prediction system.

5. Prediction Pipeline

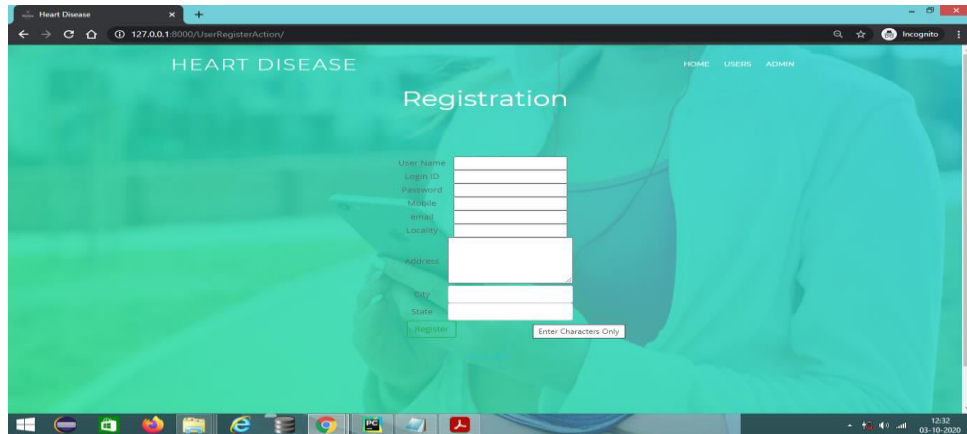
The next stage is the development of a prediction pipeline that allows new user inputs to be classified using the trained model. This pipeline takes real-time or form-based user inputs such as age, cholesterol level, and heart rate. These inputs are preprocessed in the same manner as the training data to ensure compatibility with the model. The preprocessed input is then passed into the best-performing trained model to predict whether the individual is at risk of heart disease. The result is displayed in a simple format such as "Normal" for low risk or "Abnormal" for high risk. This system can be integrated into a user interface to allow doctors or patients to easily access the prediction tool.

6. System Testing

In the final phase of the project, system testing is performed to validate the end-to-end functionality of the application. Each module — from data input to prediction output — is tested independently (unit testing) and then as a whole system (integration testing). Different types of test cases are used, including valid inputs, invalid or missing values, and edge cases to evaluate the system's robustness and error-handling capabilities. The user interface is also tested to ensure it works as intended and provides a smooth user experience. Successful testing confirms the readiness of the system for deployment or demonstration.

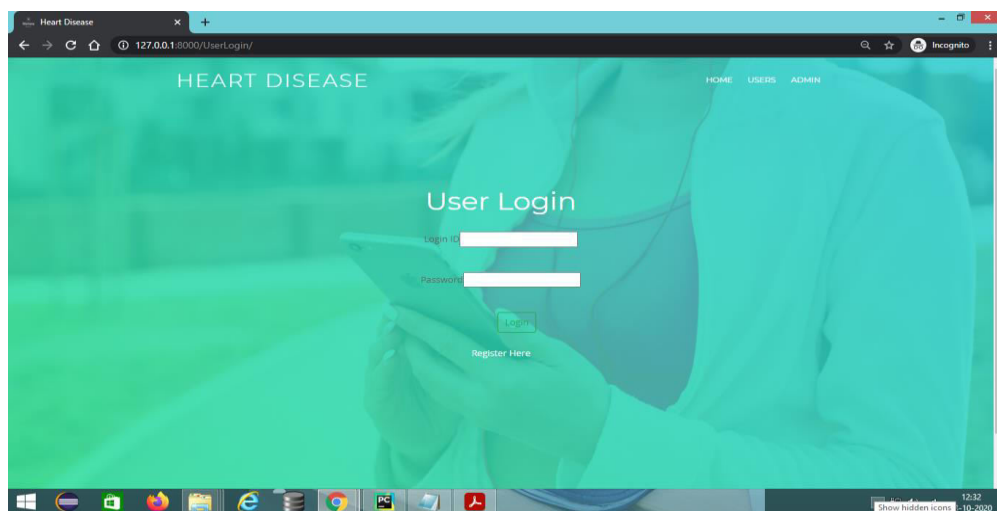


Homepage



The screenshot shows a web browser window with the title "Heart Disease". The URL bar shows "127.0.0.1:8000/UserRegisterAction/". The page has a teal background with a medical image. The title "HEART DISEASE" is at the top left, and "Registration" is in the center. There are input fields for "User Name", "Login ID", "Password", "Email", "Address", "City", "State", and "Country". A "Register" button is at the bottom. A message "Enter Characters Only" is displayed below the "Country" field. The browser's taskbar is visible at the bottom.

User Register



The screenshot shows a web browser window with the title "Heart Disease". The URL bar shows "127.0.0.1:8000/UserLogin/". The page has a teal background with a medical image. The title "HEART DISEASE" is at the top left, and "User Login" is in the center. There are input fields for "Login ID" and "Password". A "Login" button is at the bottom. A link "Register Here" is also visible. The browser's taskbar is visible at the bottom.

User Login

V. RESULT AND CONCLUSION

The evaluation phase of this project focused on comparing the performance of multiple machine learning algorithms including Support Vector Machine (SVM), Logistic Regression, Naive Bayes, K-Nearest Neighbors (KNN), and Decision Tree. Each model was trained on the same dataset and tested using unseen data to ensure fair and consistent evaluation. The comparison was based on standard performance metrics: accuracy, precision, recall, F1-score, and confusion matrix results.

Among all models, Support Vector Machine (SVM) achieved the highest accuracy, demonstrating its effectiveness in separating data points with clear boundaries. It performed well in minimizing both false positives and false negatives, making it suitable for a critical application like heart disease prediction. Logistic Regression also performed reasonably well and was particularly strong in interpretability, which is useful in medical applications where understanding the impact of features is important. The Naive Bayes model had the lowest accuracy among the tested algorithms, possibly due to its strong assumption of feature independence, which might not hold in real-world health data.

1. Security Enhancements

Security is a vital aspect of any system that deals with sensitive user data, especially in the medical domain. In this project, basic security practices were followed to ensure the protection of patient data. Input validation techniques were

implemented to prevent injection attacks, and secure storage of user input and model data was considered during the design phase. For production-level deployment, encryption protocols such as SSL for data transmission and AES for stored medical records would be recommended. Role-based access control (RBAC) can also be incorporated to restrict data access only to authorized personnel such as doctors or medical staff, maintaining the confidentiality and integrity of health information.

2. Efficiency Outcomes

The efficiency of the system is reflected through both computational performance and predictive accuracy. By implementing lightweight yet effective machine learning algorithms such as SVM and Logistic Regression, the system delivers fast prediction responses with minimal delay. Training times were kept reasonable due to dataset preprocessing and dimensionality reduction techniques. Additionally, the prediction pipeline was designed to work in real time, making the system efficient and practical for clinical or online use. The overall high accuracy and fast response time indicate that the system is optimized for performance in real-world scenarios.

3. Implementation Insights

During the implementation phase, several important insights were gained. First, proper data preprocessing plays a crucial role in the success of the models. Handling missing values, encoding categorical features correctly, and applying scaling methods helped boost model accuracy. Second, hyperparameter tuning, especially in algorithms like KNN and Decision Tree, significantly influenced model performance. Also, developing a consistent pipeline that ensures input data undergoes the same transformation during both training and prediction phases helped maintain reliability. Lastly, modular code design allowed easier testing and debugging, making the system more maintainable and scalable.

4. Advantages

This heart disease prediction system offers numerous advantages. It provides a cost-effective, quick, and automated way to assist in medical diagnosis, reducing the burden on healthcare professionals. With high predictive accuracy and user-friendly output, even non-specialists can interpret results and take early action. The system is also scalable and can be expanded to include additional health parameters or integrate with electronic health record (EHR) systems. Moreover, the modular structure allows easy updates, like adding new machine learning models or enhancing the UI. Overall, it serves as a valuable decision-support tool in early disease detection and health monitoring.

VI. FUTURE WORK

This project lays the groundwork for future enhancements that could significantly improve its scope and functionality. In the future, larger and more diverse datasets can be incorporated to improve model generalization across different populations. Integration with real-time health monitoring devices and wearable technologies could allow continuous health tracking and dynamic risk assessments. Security enhancements such as end-to-end encryption and blockchain-based medical record keeping can ensure data privacy and trust. Additionally, deep learning models like Artificial Neural Networks (ANN) or Convolutional Neural Networks (CNN) may be explored to handle more complex patterns and large datasets. Incorporating web or mobile applications for user interaction, and providing multilingual support, could expand the system's reach and usability across regions. Finally, collaboration with medical professionals to validate the model and tailor it to clinical workflows will make it more acceptable and deployable in real-world healthcare settings.

REFERENCES

1. Dua and Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019.
2. Gudadhe, Wankhade, and Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in Proc. Int. Conf. Computer and Communication Technology (IC CCT), 2010.
3. Kora and Kalva, "Improved bat algorithm for training feedforward neural networks," The Scientific World Journal, vol. 2015.
4. Mohan, Thirumalai, and Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019.
5. Zhang and Wu, "Classification of fruits using computer vision and a multiclass support vector machine," Sensors, vol. 12, no. 9, pp. 12489–12505, 2012.

6. Jadhav and Channe, "Comparative study of K-NN, naive Bayes and decision tree classification techniques," International Journal of Science and Research, vol. 5, no. 1, pp. 1842–1845, 2016.
7. Kaur and Wasan, "Empirical study on applications of data mining techniques in healthcare," Journal of Computer Science, vol. 2, no. 2, pp. 194–200, 2006.
8. Patel and Prajapati, "Study and analysis of decision tree based classification algorithms," International Journal of Computer Sciences and Engineering, vol. 6, no. 10, pp. 74–78, 2018.
9. Khashei and Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," Applied Soft Computing, vol. 10, no. 1, pp. 579–589, 2010.
10. Deo, "Machine learning in medicine," Circulation, vol. 132, no. 20, pp. 1920–1930, 2015.
11. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
12. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.
13. Witten, Frank, and Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2016.
14. Santhana Krishnan J., Geetha S. (2021). Prediction of Heart Disease using Machine Learning Algorithms. Proceedings of ICICT.
15. Aditi Gavhane, Gouthami Kokkula, Isha Panday, Kailash Devadkar (2020). Prediction of Heart Disease using Machine Learning. 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA).
16. Senthil Kumar Mohan, Chandrasegar Thirumalai, Gautam Srivastva (2022). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.
17. Himanshu Sharma, M. A. Rizvi (2023). Prediction of Heart Disease using Machine Learning Algorithms: A Survey. International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), Vol. 5.
18. M. Nikhil Kumar, K. V. S. Koushik, K. Deepak (2022). Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology (IJSRCSEIT).
19. Pahulpreet Singh Kohli, Shriya Arora (2021). Application of Machine Learning in Disease Prediction. 4th International Conference on Computing, Communication, and Automation (ICCCA).
20. M. Akhil, B. L. Deekshatulu, P. Chandra (2020). Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm. Procedia Technology, Vol. 10, pp. 85–94.
21. Hazra A., Mandal S., Gupta A., Mukherjee (2022). Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology.
22. Himanshu Sharma, M. A. Rizvi (2021). Prediction of Heart Disease using Machine Learning Algorithms: A Survey. International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), Vol. 5.
23. Patel J., Upadhyay P., Patel (2022). Heart Disease Prediction Using Machine Learning and Data Mining Techniques. Journal of Computer Science & Electronics.
24. Pahulpreet Singh Kohli, Shriya Arora (2021). Application of Machine Learning in Disease Prediction. 4th International Conference on Computing, Communication, and Automation (ICCCA).
25. Pahulpreet Singh Kohli, Shriya Arora (2021). Application of Machine Learning in Disease Prediction. 4th International Conference on Computing, Communication, and Automation (ICCCA).



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details